Is an apple like a fruit?

Is an apple like a fruit?
A study on comparison and categorisation statements

Paula Rubio-Fernández

University College London

Bart Geurts

University of Nijmegen

Chris Cummins

University of Edinburgh

Abstract

Categorisation models of metaphor interpretation are based on the premiss that categorisation statements (e.g., 'Wilma is a nurse') and comparison statements (e.g., 'Betty is like a nurse') are fundamentally different types of assertion. Against this assumption, we argue that the difference is merely a quantitative one: '*x* is a *y*' unilaterally entails '*x* is like a *y*', and therefore the latter is merely weaker than the former. Moreover, if '*x* is like a *y*' licenses the inference that *x* is not a *y*, then that inference is a scalar implicature. We defend these claims partly on theoretical grounds and partly on the basis of experimental evidence. A suite of experiments indicates both that '*x* is a *y*' unilaterally entails that *x* is like a *y*, and that in several respects the non-*y* inference behaves exactly as one should expect from a scalar implicature. We discuss the implications of our view of categorisation and comparison statements for categorisation models of metaphor interpretation.

Is an apple like a fruit?

A study on comparison and categorisation statements

Abstract

Categorisation models of metaphor interpretation are based on the premiss that categorisation statements (e.g., 'Wilma is a nurse') and comparison statements (e.g., 'Betty is like a nurse') are fundamentally different types of assertion. Against this assumption, we argue that the difference is merely a quantitative one: '*x* is a *y*' unilaterally entails '*x* is like a *y*', and therefore the latter is merely weaker than the former. Moreover, if '*x* is like a *y*' licenses the inference that *x* is not a *y*, then that inference is a scalar implicature. We defend these claims partly on theoretical grounds and partly on the basis of experimental evidence. A suite of experiments indicates both that '*x* is a *y*' unilaterally entails that *x* is like a *y*, and that in several respects the non-*y* inference behaves exactly as one should expect from a scalar implicature. We discuss the implications of our view of categorisation and comparison statements for categorisation models of metaphor interpretation.

1. Introduction

The distinction between comparison and categorisation statements has been of central importance to theoretical debates on the metaphor-simile distinction – a recurrent theme in the metaphor literature for the last forty years (e.g., Kintsch, 1974; Ortony, 1979; Glucksberg & Keysar, 1990; Kennedy & Chiappe, 1999; Carston, 2002; Glucksberg,

2008; Carston & Wearing, 2011; Barnden, 2012).[1] This distinction has also been central to the numerous empirical studies that have compared the use and interpretation of metaphors and similes (e.g., Johnson, 1996; Chiappe & Kennedy, 1999; Chiappe et al., 2003a; Glucksberg & Haught, 2006a; Utsumi, 2007; Pierce & Chiappe, 2008; Haught, 2013a). Generally speaking, standard models in the Aristotelian tradition have emphasised the commonalities between metaphors and similes, while more recent accounts underline the differences between the two.

To illustrate what is at issue, consider the following examples:

(1) A papaya is a fruit.
(2) ?A papaya is like a fruit.
(3) A papaya is like a mango.
(4) My lawyer is a shark.
(5) My lawyer is like a shark.

As the last two examples illustrate, metaphors and their corresponding similes can be used interchangeably. However, turning a literal categorisation statement into a comparison statement is not unproblematic, as the first two examples show. This is what Glucksberg (2011) calls 'the paradox of unlike things compared'. On the traditional view, metaphors are implicit similes (see, e.g., Aristotle in Barnes, 1984; Kintsch, 1974; Grice, 1975; Miller, 1979; Ortony, 1979; Searle, 1979; Fogelin, 1988; Gentner et al., 2001; Israel et al., 2005). This is why one type of sentence can be used to paraphrase the other, as in (4) and (5). By contrast, categorisation models of metaphor interpretation claim that metaphors and similes are understood in fundamentally different ways: 'the metaphor as a categorisation assertion, the simile as an assertion of

[1] Following this literature, we understand similes as figurative comparisons (e.g., 'My love is like a rose'), which are different from literal comparisons (e.g., 'My love is like her mother'). Seen this way, similes are the comparison counterpart of nominal metaphors (e.g., 'My love is a rose'), which by definition involve a category violation.

similitude' (Glucksberg & Haught, 2006a:361; see also Chiappe & Kennedy, 2001; Glucksberg, 2001, 2011; Carston, 2002; Carston & Wearing, 2011; Haught, 2013a). Against the traditional view, categorisation models of metaphor interpretation have often stressed how the interpretation of metaphors and similes differ, despite their apparent interchangeability (see Glucksberg & Haught, 2006a, 2006b and Haught, 2013a, 2013b, for empirical evidence).

According to the categorisation view of metaphor, the reason why it is possible to paraphrase a simile as a nominal metaphor and vice versa is that in the simile the 'vehicle' of the figurative expression (e.g., the word 'shark' in (4) and (5)) stands for the literal concept, whereas in the metaphor the vehicle stands for a superordinate category including not only members of the literal category (e.g., real sharks) but also other entities that share relevant properties with the prototype for the superordinate category (e.g., my lawyer). The fact that metaphor vehicles are polysemous between their literal and figurative meanings explains that the same word can be used to express both a comparison statement (as in example (5)) and a categorisation statement (as in (4)) , with the same word naming distinct categories at different levels of abstraction (i.e. the literal concept SHARK in the simile and the superordinate category SHARK* in the metaphor;[2] see, e.g., Glucksberg & Keysar, 1990; Glucksberg, 2001, 2011; Carston, 2002; Glucksberg & Haught, 2006a; Sperber & Wilson, 2008; Haught, 2013a).[3]

[2] It is worth noting that not only metaphors but also brand names can become polysemous in this way. For example, the name 'Kleenex' can be used to refer to a specific type of paper tissue or to paper tissues in general (see Glucksberg & Keysar, 1990 and Sperber & Wilson, 2008, for a discussion of this and other examples). It is the contention of categorisation models that this kind of language use is ubiquitous in everyday language (including sign languages) and therefore does not require a special interpretation mechanism that distinguishes metaphorical language from literal language.

[3] See also Glucksberg & Keysar (1990, 1993) for an earlier categorisation model of metaphor interpretation, according to which similes are implicit metaphors (i.e. comparisons to a superordinate category; cf. Carston, 2002).

In summary, according to the comparison view, examples (4) and (5) are equivalent because metaphors are implicit similes. By contrast, according to categorisation models, the nominal metaphor in (4) is the figurative counterpart of the literal categorisation statement in (1), while the simile in (5) is the figurative counterpart of the literal comparison statement in (3).

In the present study we investigated the comprehension of literal comparison and categorisation statements. Our starting point was the distinction drawn by categorisation models of metaphor interpretation between similes and metaphors as fundamentally distinct types of expression. In defending this view, Keysar makes the following argument:

> Category membership is incompatible with assertions of similarity, e.g., if 'Copper is a metal' is acceptable, then 'Copper is *like* a metal' is not acceptable. Similarly, if one asserts and believes that someone is an actual baby, then one cannot simultaneously assume that the person is 'like' a baby. (Keysar, 1989: 380-381)

The same argument has been made on the basis of similar examples in various other papers defending the categorisation view of metaphor (see, e.g., Glucksberg & Keysar, 1990, 1993; Kennedy & Chiappe, 1999; Chiappe & Kennedy, 2001; Carston, 2002; Glucksberg, 2003; Glucksberg & Haught, 2006a; Barnden, 2012). In a similar vein, Glucksberg makes the following generalisation:

> The relative position of terms within such hierarchical categories determines when comparisons are permissible and when categorical assertions are permissible. In general, comparisons are restricted to terms that refer at the same level of abstraction. Thus, we can have comparisons between superordinates, as in *fresh fruits are like salad greens*, but not

> between superordinates and subordinates within a category, as in *lettuce is like salad greens* or *romaine lettuce is like lettuce*. When two entities are at different levels in a taxonomic hierarchy, then the appropriate relation is categorical, not one of similitude, as in *lettuce is a salad green* or *romain is a (kind of) lettuce*. (Glucksberg, 2001: 42)

While we agree with the view that interpreting novel nominal metaphors involves the construction of a superordinate category on the basis of the metaphor vehicle (Rubio-Fernández, 2007), we contest the claim that a comparison statement ('*x* is like a *y*') contradicts the corresponding categorisation statement ('*x* is a *y*'). We argue (A) that categorisation statements are stronger than, and therefore compatible with, comparison statements,[4] and (B) that *if* an utterance of '*x* is a *y*' licenses the inference that *x* not a *y*, this inference is pragmatic in nature; specifically, it is a scalar implicature. Therefore, should statements like 'An apple is like a fruit' turn out to be infelicitous (which remains to be seen), then the cause is pragmatic rather than semantic.

The aim of the present study is to provide empirical evidence for these hypotheses. The paper is structured as follows. Having developed in some detail our Hypotheses A and B (Section 2), we present a series of Mechanical-Turk experiments we conducted to test our hypotheses.[5] (For better readability, this discussion will skimp on methodological and statistical details, which are presented in full in the Appendix.) Section 3 gives evidence that '*x* is a *y*' is semantically stronger than '*x* is like a *y*'; Section 4 shows that, at least some of the time, utterances of '*x* is like a *y*' license the inference that *x* is not a *y*;

[4] A related argument has been made by Kennedy and Chiappe (2001) who argued that literal categorisation statements (e.g., 'That is an apple') are used when the two concepts share many common properties, while comparison statements (e.g., 'That is like an apple') are used when they share few common properties.

[5] It has been shown that the quality of data gathered through Amazon's Mechanical Turk is comparable to that of laboratory data (Schnoebelen & Kuperman, 2010; Buhrmester et al., 2011; Sprouse, 2011; Crump et al., 2013).

and Section 5 presents data that comport with our hypothesis that this inference is a scalar implicature. Finally, in Section 6, we discuss the implications of our account for categorisation models of metaphor interpretation.

Before we get started, we should make it clear that the hypotheses we are about to defend do not amount to a fully fledged account of categorisation and comparison statements, let alone a theory of metaphor and simile comprehension. We do believe, however, that our hypotheses, if true, impose substantial constraints on theories that attempt to deal with the metaphor-simile distinction and have debated it for decades. Hence, we will discuss some of the implications of our results for categorisation models of metaphor interpretation at the end of the paper, though will do so only at a general level; a detailed theoretical analysis is beyond the scope of this paper.

2. *Unilateral entailment* and *scalar implicature*

From a semantical point of view, to say that '*x* is a *y*' is stronger than '*x* is like a *y*' is to say that, by virtue of their respective meanings, the former sentence is true whenever latter is true, but not vice versa. The key notion here is 'entailment', which is standardly defined as follows (where S_1 and S_2 are arbitrary sentences):

> *Entailment*
>
> S_1 entails S_2 if and only if S_2 is true whenever S_1 is true.

For example, (6) and (7) each entail (8) and they entail one another too: (6) is true whenever (7) is true and vice versa.

(6) Fred is an oculist.

(7) Fred is an ophthalmologist.

(8) Fred is a physician.

If two sentences entail one another, they are equally strong. If one sentence entails another but not the other way round, the former makes a stronger a statement than the latter. This is unilateral entailment:

> *Unilateral entailment*
> S_1 unilaterally entails S_2 if and only if S_1 entails S_2 but not the other way round.

For example, (6) and (7) each unilaterally entail (8) but not the other. Hence, (6) makes a stronger statement than (8), and so does (7).

Thus, the first hypothesis we will defend in this paper is the following:

(A) '*x* is a *y*' unilaterally entails '*x* is like a *y*'.

It is a recurrent observation in the metaphor literature that nominal metaphors are 'stronger', more 'direct' or more 'forceful' than the corresponding similes (e.g., Ortony, 1979; Fogelin,1988; Glucksberg & Keysar, 1993; Stern, 2000; Carston, 2002; Zharikov & Gentner, 2002; Chiappe et al., 2003a; Israel et al., 2005; Glucksberg, 2011; cf. O'Donoghue, 2009). The most straightforward explanation for these intuitions is that '*x* is a *y*' is stronger than '*x* is like a *y*' simply by virtue of what these sentences mean; that is to say, the reason why '*x* is a *y*' appears stronger than '*x* is like a *y*' is that the former unilaterally entails the latter.

Another intuition that is often voiced in the literature is that '*x* is a *y*' is incompatible with '*x* is like a *y*'. A possible reason for this incompatibility is that '*x* is like a *y*' implies that *x* is not a *y*. If this intuition is correct, it would seem to contradict our Hypothesis A. For if '*x* is a *y*' implies '*x* is like a *y*', as Hypothesis A has it, and '*x* is like a *y*' implies that *x* is not a *y*, then by transitivity '*x* is a *y*' implies that *x* is not a *y*, which is clearly contradictory.

This paradox is resolved by our second hypothesis:

(B) If an utterance of '*x* is like a *y*' licenses the inference that *x* is not a *y*, then this inference is a scalar implicature.

To explain the notion of scalar implicature, consider the following examples (see Geurts, 2010 for a review):

(9) Fred ate all the cookies.

(10) Fred ate some of the cookies.

(9) is clearly stronger than (10): if Fred ate all the cookies, he must have eaten some of the cookies, but the converse doesn't hold. Nonetheless, it is a well-known observation that an utterance of (10) may license the inference that, according to the speaker, Fred did not eat all the cookies and therefore (9) is false. This inference can be accounted for as a pragmatic inference; that is to say, as an inference that follows, not from the sentence as such, but from the fact that a speaker utters this sentence in a given context and that the hearer is entitled to reason as follows:

Scalar implicature

If the speaker believed that (10) is true, he should have said so, but instead he chose to make a weaker claim, (9). Therefore, he probably doesn't believe that (10) is true.

Sentence (9) unilaterally entails sentence (10), and *by uttering* sentence (10) a speaker may license the implicature that he doesn't accept that (9) is the case. But it clearly doesn't follow that sentence (9) implies its own falsehood. Likewise, '*x* is a *y*' unilaterally entails '*x* is like a *y*' (Hypothesis A), and by uttering '*x* is like a *y*' a speaker may license the implicature that he doesn't accept that *x* is a *y* (Hypothesis B). But it doesn't follow that '*x* is a *y*' implies its own falsehood.

It follows from our Hypothesis A that, contrary to what has been claimed by proponents of the categorisation theory of metaphor, corresponding categorisation and comparison statements are not incompatible with one another; for according to Hypothesis A, if '*x* is a *y*' is true, then '*x* is like a *y*' is true as well. Nonetheless, there appears to be a rather strong intuition that '*x* is like a *y*' implies that *x* is not a *y*, and Hypothesis B explains where this intuition comes from: it is a pragmatic inference, namely a scalar implicature. Therefore, the tension between '*x* is a *y*' and '*x* is like a *y*' doesn't run as deep as categorisation theories of metaphor have claimed: the two types of statement are compatible at the semantical level; the tension is merely pragmatic.[6]

3. Evidence for Hypothesis A: '*x* is *y*' unilaterally entails '*x* is like a *y*'

That categorisation statements are stronger than comparison statements is indicated by the standard diagnostics (Horn, 1989; Matsumoto, 1995).[7]

---

[6] According to our reviewers, sentences like (i) are problematic for our account:

(i) Nixon was a Quaker, but he was not like a Quaker.

(For the benefit of any readers under 40 that may have wandered into this article: Richard Milhous Nixon was the 37th president of the United States, and he was in fact a Quaker.) Assuming that (i) is felicitous at all (and we are not entirely convinced that it is; cf. 'Nixon was a Quaker, but he did not behave like a Quaker'), it may seem to defy our hypothesis that 'Nixon was a Quaker' entails 'Nixon was like a Quaker', because the entailment would render (i) contradictory. But rather than giving rise to a contradiction, (i) seems to imply that Nixon was not like an ordinary or 'normal' Quaker. This suggests that the two occurrences of 'a Quaker' denote slightly different concepts, which is not uncommon in contrastive environments (Geurts, 1998), as (ii) illustrates:

(ii) That's not a car, it's a Ferrari.

This kind of contrast appears to be required in order to support the intended meaning of (i), which we take to be that Nixon was by definition a Quaker, but was unlike an archetypal Quaker.

[7] Note that all these examples are marked in the linguistic sense of the word, and therefore would typically be used in somewhat special circumstances; for example, to correct another speaker. However, this doesn't affect our argument in any way (see Chiappe & Kennedy, 2000 for a discussion of the use of metaphors to correct similes).

(11) A kumquat is like a citrus fruit, and in fact it IS a citrus fruit.

(12) Not only is a kumquat LIKE a citrus fruit, it IS a citrus fruit.

(13) I'm pretty sure that a kumquat is like a citrus fruit, and for all I know it might even BE a citrus fruit.

The mere fact that these sentences are felicitous shows that '*x* is like a *y*' and '*x* is a *y*' are compatible. Furthermore, each of these constructions can only be used if the first statement is weaker than the second. Cf.

(14) She is intelligent, and in fact she is brilliant.

(15) ?He is intelligent, and in fact he is tall.

These observations already indicate that '*x* is a *y*' is merely stronger than '*x* is like a *y*', and they are corroborated by web data:

(16) It's like a registry key name (and, in fact, on PC it is a registry key).

(17) She is like a sister to me; in fact, she is my sister from another mother.

Further evidence is provided by the following observation. The inference pattern 'If S_1 then S_2; therefore if S_3 then S_2' only is valid if S_3 entails S_1. Now consider the following pair of sentences:

(18) If kumquats are like citrus fruits, we can use them for our Christmas punch.

(19) If kumquats are citrus fruits, we can use them for our Christmas punch.

Clearly, a speaker who accepts (18) is committed to the truth of (19), so again it follows that a categorisation statement entails the corresponding comparison statement.

Experiment 1a investigated how people would interpret comparison and categorisation statements in a sentence verification task using made-up words accompanied by summary definitions consisting of three attributes; Figure 1 shows a sample item. The

point of the first experiment was to investigate people's interpretations of comparison and categorization statements independently from their world knowledge.

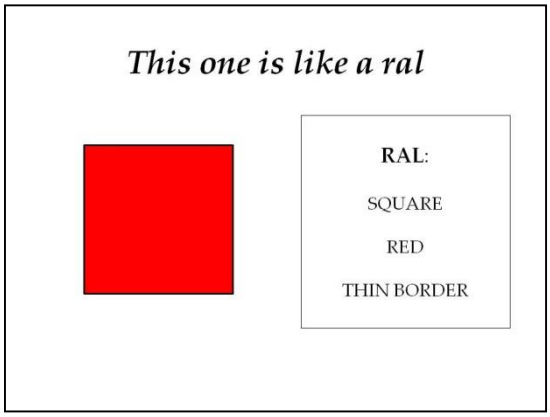


Figure 1: Item from Experiment 1a (the figure satisfies all the properties in the definition).

The shapes on display had between zero and three of the attributes included in the definition, so there were four levels of similarity between the figure and the definition, which we will refer to as L_0-L_3. The target sentences were comparison and categorisation statements (e.g., 'This one is like a ral' vs. 'This one is a ral'). In each case, participants had to indicate whether the sentence was true or false of the shape on display. As shown in Figure 2, categorisation statements were accepted only in the L_3-condition, and in this condition the average rate of 'true' responses was close to 100%; in the L_0, L_1, and L_2 conditions, the corresponding rates were below 5%. The comparison statements, too, were close to ceiling level in the L_3-condition, but their rates of 'true' responses didn't fall as abruptly as with the categorisation statements; rather, they declined gradually, and only the L_0-items were unanimously rejected. This pattern was replicated in Experiment 1b, which used the same design as Experiment 1a, except that participants only saw comparison statements (see the Appendix for further details).

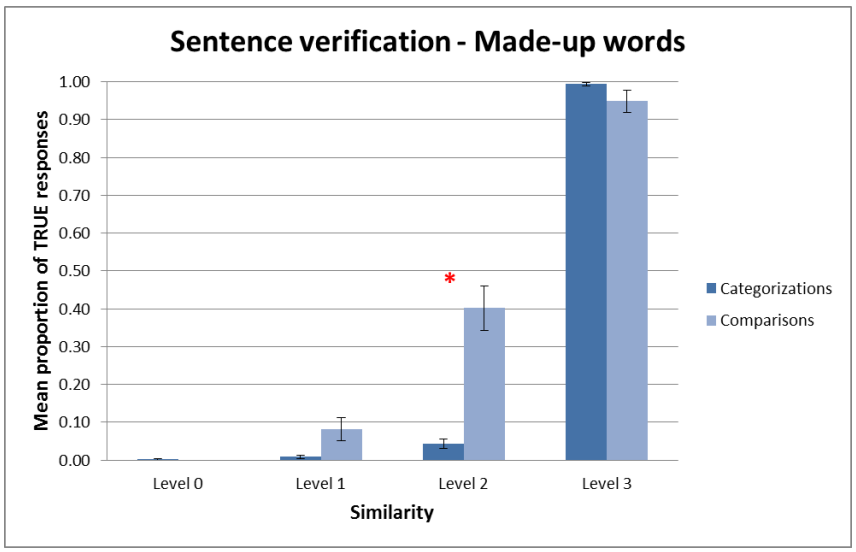


Figure 2: Mean proportions of TRUE responses to categorisation and comparisons statements in Experiment 1a (SE bars; asterisk $p < .001$).

Our results are clearly at odds with the claim that a comparison statement is incompatible with the corresponding categorisation statement. Secondly, the response pattern observed in Experiment 1a shows that, in the context of this experiment, at least, '*x* is a *y*' unilaterally entails '*x* is like a *y*'; for the latter was accepted whenever the former was, and not vice versa. This is in line with Hypothesis A, and contradicts the claim that categorisation and comparison statements are semantically incompatible.

Since the materials used in Experiments 1a and 1b were patently artificial, we replicated Experiment 1a using a familiar category, namely animals; Figure 3 shows a sample item. Participants saw pictures of a tiger, for example, accompanied by either a categorisation statement ('This one is a tiger') or a comparison statement ('This one is like a tiger'). The predicate denoted either a superordinate ('wild animal' for tiger), the same category as the referent ('tiger'), a merely similar category ('lion') or a dissimilar one ('bear').

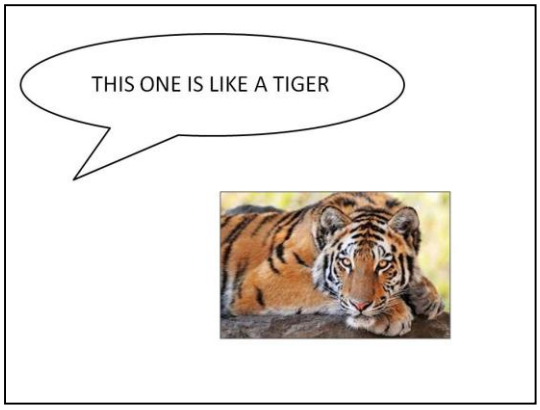
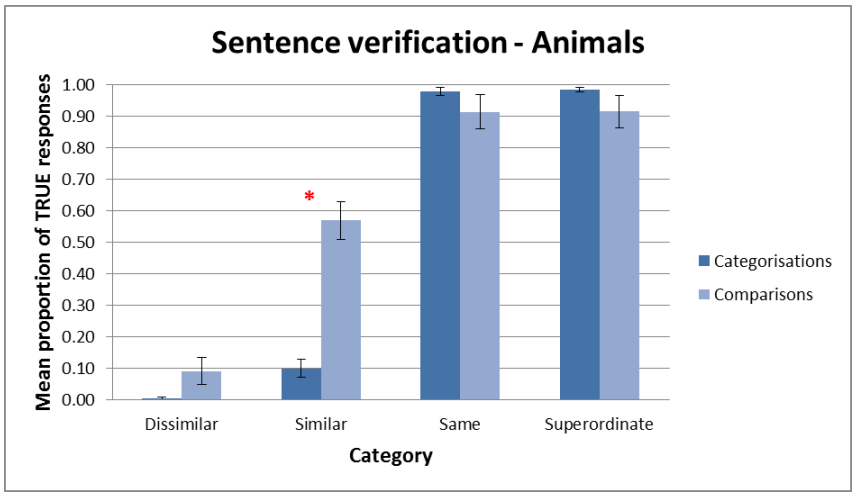


Figure 3: Item from Experiment 2.

The results of Experiment 2 were virtually identical to those of the first experiments (see Figure 4). Most importantly, comparison statements of the form 'This one is like a *y*' were accepted over 90% of the time if the picture showed a *y*.[8] Hence, the conclusions we drew from the first experiments remain unchanged: we find no support for claims to the effect that '*x* is like a *y*' and '*x* is a *y*' are incompatible, and it appears that according to impartial informants '*x* is a *y*' unilaterally entails '*x* is like a *y*', since the latter was accepted whenever the former was, but not vice versa.



[8] Materials were presented in a quasi-random order so that participants had to verify 'This one is a tiger' before the corresponding comparison statement. Thus we ensured that participants had correctly identified the animal in question, and didn't agree to 'This one is like a tiger' because they thought it was a similar animal, though not a tiger.

Figure 4: Mean proportions of TRUE responses to categorisation and comparison statements in Experiment 2 (SE bars; asterisk $p < .001$).

4. Evidence that '*x* is like a *y*' licenses the inference that *x* is not a *y* (at least some of the time)

In Experiments 1a, 1b and 2 we observed hardly any instances in which participants took '*x* is like a *y*' to imply that *x* is not a *y*. This finding is consistent with our Hypothesis B, which states that *if* a comparison statement licenses this inference, it is due to a scalar implicature and not an entailment. However, our results sit somewhat uneasily with intuitions reported by Glucksberg and Keysar, Carston, Chiappe and others to the effect that, for example, 'A nuthatch is like a bird' is an unacceptable statement. Therefore, we conducted another experiment, which aimed to determine whether the non-*y* inference would emerge in a different set-up.

In Experiment 3 we adopted an inference paradigm, asking participants directly whether they would draw a non-*y* inference from a statement of the form '*x* is like a *y*', using trials like the following:

> John says: 'My mother is like a nurse.'
> Would you conclude from this that, according to John, his mother is not a nurse?

Generally speaking, inference tasks of this type yield higher proportions of positive results than the corresponding verification tasks (Evans et al., 1993; Geurts & Pouscoulous, 2009; Geurts, 2010). There are various possible reasons for this, which need not exclude each other, but the most obvious one is that people are generally more likely to accept an inference that is presented explicitly. We therefore expected higher levels of non-*y* inferences in Experiment 3 than in Experiments 1a, 1b and 2.

	CONDITIONS		SAMPLE ITEMS	
	Inference	*Knowledge*	*Statement*	*Conclusion*
(A)	True	Common	A zebra is like a horse	A zebra is not a horse
(B)	False	Common	A robin is like a bird	A robin is not a bird
(C)	True (lit.)	Common	The coach's voice is like a foghorn	The coach's voice isn't a foghorn
(D)	Possible	Private	My watch is like a Rolex	My watch is not a Rolex

Table 1: Conditions and sample items from Experiment 3 (inference task).

Table 1 summarises the four conditions used in Experiment 3. As can be seen in this table, three of the conditions relied on common knowledge: in condition (A), it was common knowledge that the target inference is true; in condition (B), that it is false; and in condition (C), that is true but only if the predicate is construed literally. Condition (D) relied on the speaker's private knowledge.

We expected condition (A) to reveal the highest proportion of positive responses, because it should benefit from a positive belief bias: participants were likely to know that the non-*y* inference was true, and thus be biased towards a positive response. Contrariwise, we expected condition (B) to yield the lowest agreement rates since participants were likely to know that the non-*y* inference was false, and thus be biased towards a negative response (i.e. negative belief bias). Thirdly, we expected the results of condition (D) to be between those in conditions (A) and (B) since the non-*y* inference relied on the speaker's private knowledge and hence participants shouldn't suffer from any kind of belief bias. Finally, we expected the results of condition (C) to be intermediate, too, since positive responses depended on a literal construal of the conclusion. That is, if taken literally, the C-inferences were obviously true since *x* and *y*

were unrelated concepts in that condition. However, given the figurative interpretation of the C-statements (which were effectively similes), participants may be prompted to interpret the conclusion as a negated metaphor and hence reject it on the basis that if a simile is true, then the corresponding metaphor must also be true.

The results of Experiment 3 are shown in Figure 5. As predicted, A-inferences were endorsed significantly more often than all others; B-inferences showed the lowest agreement rates, but were nonetheless endorsed more than half the time (67%), despite the negative belief bias;[9] and the rates for C- and D-inferences were in the mid-range and statistically indistinguishable.

A pattern of results that is relevant for theories of metaphor interpretation is the difference observed between conditions A and C: while the conclusion was true in both of these conditions (e.g., 'A zebra is not a horse' and 'The coach's voice is not a foghorn'), participants were more prone to endorse the inference in condition A than in condition C. We interpret these results as evidence that in the latter condition, participants sometimes interpreted the negative conclusion as a negated metaphor (which they rejected because the corresponding simile was stated as true). That similes and metaphors were interchangeable in this task was confirmed in a control condition, in which participants agreed to the simile version of a metaphor 84% of the time.

Experiment 3 shows that comparison statements of the form '*x* like a *y*' give rise to non-*y* inferences at least some of the time. The following section presents evidence that these non-*y* inferences are scalar implicatures.

[9] As pointed out to us by one of the reviewers for this journal, the fact that agreement rates in the B-condition were quite high despite the negative belief bias is additional evidence for our analysis. See Geurts (2010: 157-158) for discussion.

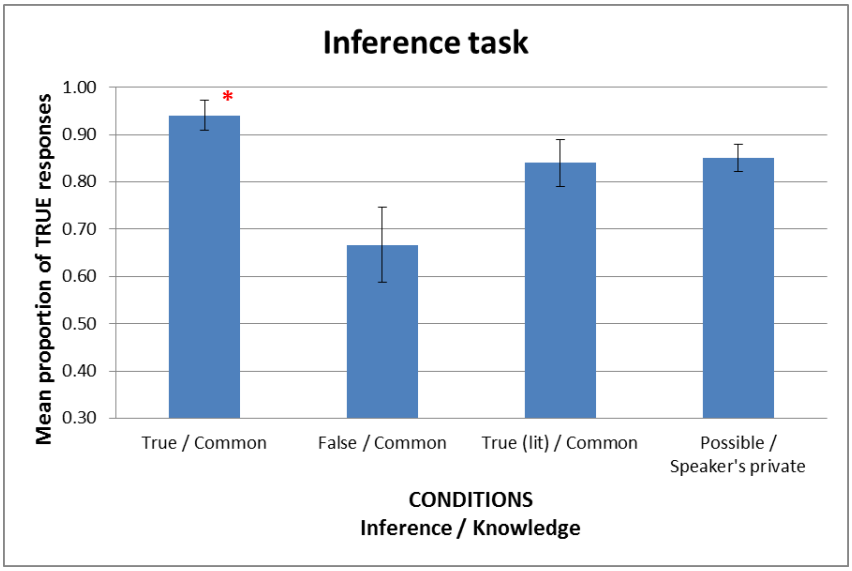


Figure 5: Mean proportions of TRUE responses to the four types of conclusions used in Experiment 3 (SE bars; asterisk $p < .083$).

5. Evidence for Hypothesis B: the non-*y* inference is a scalar implicature

In order to explain the argument of this section, let us have a closer look at a run-of-the-mill example of scalar implicature:

(20) Some of the apples are red.
(21) All the apples are red.

An utterance of (20) may give rise to the implicature that, according to the speaker, (21) is false, i.e. that not all the apples are red. This inference is based on two assumptions that we left implicit so far:

1. It is presupposed that, if the speaker knows (or believes) that all the apples are red, then he should have said (21) rather than (20), even if strictly speaking both sentences are true. That is to say, between these two sentences, (21) is the preferred way of describing a situation in which all the apples are red.

2. It is presupposed that the speaker knows whether or not all the apples are red. If the speaker isn't knowledgeable with respect to the stronger claim (i.e. whether or not all the apples are red), the hearer is not entitled to infer that, according to the speaker, the stronger claim is false. The assumption that the speaker is knowledgeable with respect to the stronger claim is often called the 'competence assumption' (Soames, 1982; Horn, 1989; Sauerland, 2004; van Rooij & Schulz, 2004; see Geurts, 2010 for discussion and Goodman & Stuhlmüller, 2013 for experimental evidence).

In conjunction with our Hypothesis B, these observations yield the following predictions:

1. If a categorisation statement is true, it should be preferred to the corresponding comparison statement; that is to say, if *x* is a *y*, then '*x* is a *y*' should be preferred to '*x* is like a *y*'.[10] We tested this prediction in Experiment 4.
2. If there is reason to suppose that the speaker doesn't know whether or not *x* is a *y*, then this should affect the likelihood that his uttering '*x* is like a *y*' is felt to imply that *x* is not a *y*. This prediction was tested in Experiment 5.

The materials used in Experiment 4 were similar materials to those used in Experiment 2, but instead of asking people whether a certain description was true or false of a depicted animal, we asked them to rate descriptions of animals, vegetables, and fruits on a scale from 1 ('completely unacceptable') to 7 ('perfectly acceptable').[11] Figure 6 shows a sample item. As in Experiments 1a and 2, the materials included equal

[10] Note that the preference for the stronger statement (if true) is a prerequisite for, but not the same thing as, the implicature that the stronger statement is false. Katsos & Bishop (2011) argue, correctly in our view, that some experimental studies have mixed these two things up, and show that they are dissociated in 5- and 6-year-old children. Relatedly, in order to explain why '*x* is like a *y*' is infelicitous if *x* is known to be a *y*, we need not assume that a not-*y* implicature is derived.

[11] For extensive discussion on the use of rating tasks in implicature studies, see Geurts & van Tiel (2013) and Van Tiel (2014a, b).

numbers of categorisation and comparison statements. Orthogonally to this division, three types of sentences were used: (*a*) sentences with basic-level predicates (e.g., 'Labradors are (like) dogs'), (*b*) sentences with superordinate predicates (e.g., 'Sharks are (like) predators'), and (*c*) sentences in which the head noun of the subject reappeared in the predicate (e.g., 'Grizzly bears are (like) bears'). Finally, the materials included a control condition in the form of unobjectionable comparison statements (e.g., 'Wild boars are like pigs').

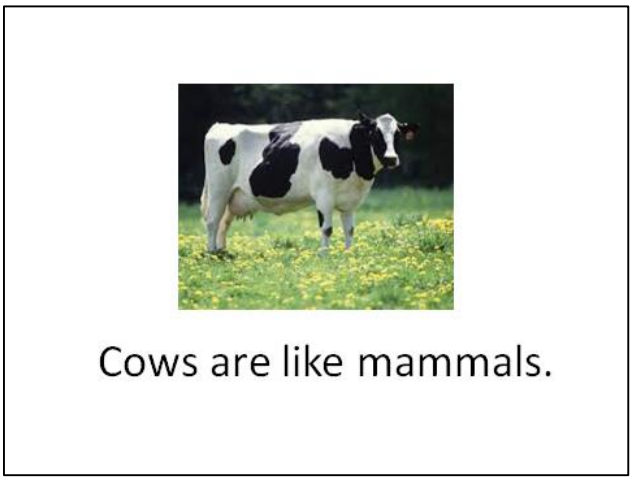


Figure 6: Sample item from Experiment 4 (rating task

).

We predict that if *x* is a *y*, the sentence '*x* is like a *y*' should receive a poorer rating than '*x* is a *y*'. As can be seen in Figure 7, the results of Experiment 4 confirmed this prediction. Whereas categorisation statements attained mean ratings at the top end of the scale, the ratings for comparison statements remained close to midpoint across the board, and in each condition ranked significantly lower than categorisation statements and felicitous controls of the 'Wild boars are like pigs' type. These findings agree with our first prediction.

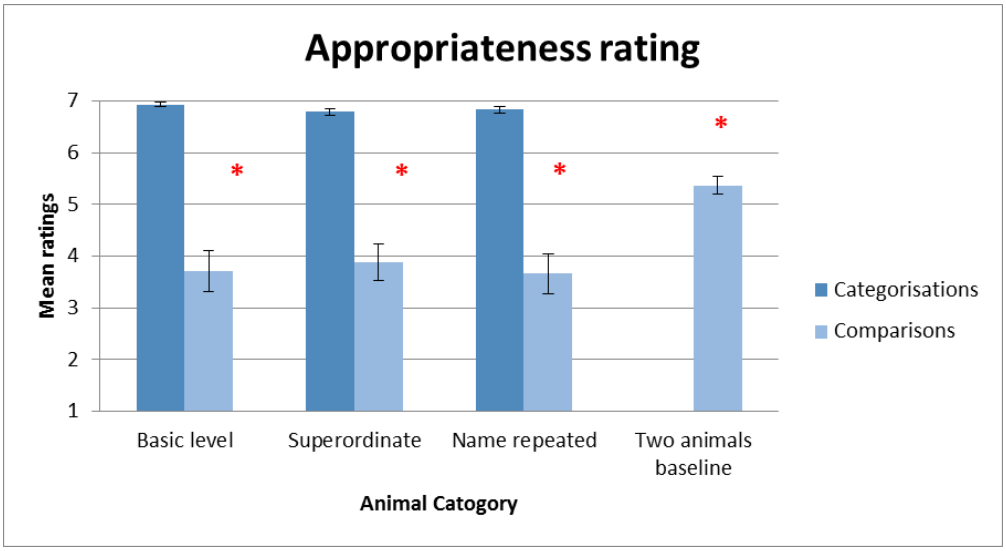


Figure 7: Mean appropriateness rates for categorisation and comparison statements in Experiment 4 (SE bars; asterisk $p < .002$).

Our last experiment, number 5, was inspired by Goodman & Stuhlmüller's (2013) study. Participants were presented with the following vignette:

> The Municipal Museum has a painting called *The Swedish Horseman*, which was thought to be of no great value. However, a visiting expert expressed the opinion that it might actually be by Rembrandt. The museum director didn't believe this, but agreed to submit the painting for testing.
>
> The painting is to be subjected to a series of 9 tests. It is agreed that if all the tests come back positive, there can be no doubt that the *The Swedish Horseman* was painted by Rembrandt. However, if one or more of the tests comes back negative, the painting is not a Rembrandt.

Participants were evenly and randomly allocated to one out of four conditions, two of which were critical, while the other two were controls. In the critical conditions, the story continued in either of the following ways:

> As of today, (*a*) all of the tests // (*b*) 7 of the 9 tests have been completed. A reporter asks the museum director to comment on the painting, and the director replies that:
>
> > *The Swedish Horseman* is like a Rembrandt.
>
> Would you conclude from this statement that, according to the director, *The Swedish Horseman* is *not* a genuine Rembrandt?

In the (*a*) version, the competence assumption was justified: it was reasonable to suppose that the director knew whether or not *The Swedish Horseman* is a Rembrandt. In the (*b*) version, this assumption was not justified.

In the control conditions, the story continued in either of the following ways:

> As of today, (*a*) all of the tests // (*b*) 7 of the 9 tests have been completed. A reporter asks the museum director to comment on the painting, and the director replies that:
>
> > Most of the tests came back positive.
>
> Would you conclude from this statement that, according to the director, some of the tests came back negative?

In the (*a*) version, the competence assumption was justified: it was reasonable to suppose that the director knew whether or not all the tests came back positive. In the (*b*) version, this assumption was not justified.

Participants were given a 1-7 Likert scale to respond, where 7 was defined to mean that they definitely agreed with the suggested conclusion, while 1 was defined to mean that they definitely disagreed. Our prediction was that, in both the critical and the control conditions, ratings would be higher for the (*a*) version than for the (*b*) version. As shown in Figure 8, this prediction was borne out by the data.

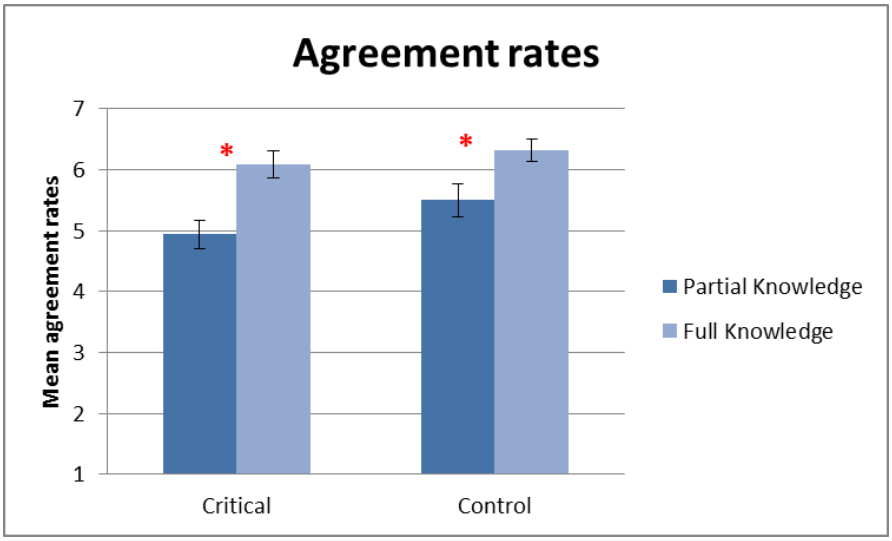


Figure 8: Mean agreement rates for the comparison statements (critical condition) and the quantified statements (control condition) used in Experiment 5 as a measure of the effect of the speaker's competence in the derivation of scalar implicatures (SE bars; asterisk $p < .02$).

The evidence presented in this section is consistent with two predictions which follow from our hypothesis that the non-*y* inference associated with '*x* is like a *y*' statements are scalar implicatures. Our first prediction was that if *x* is a *y*, then '*x* is a *y*' should be preferred to '*x* is like a *y*'; Experiment 4 indicates that such a preference does exist. Our second prediction was that the speaker's knowledge as to whether or not *x* is a *y* should affect the likelihood that his uttering '*x* is like a *y*' is felt to imply that *x* is not a *y*; Experiment 5 provides evidence that this prediction is correct.

6. Discussion and conclusions

Both our theoretical arguments and empirical evidence were developed on the basis of literal language use. How is this study then relevant to models of metaphor interpretation? Importantly, those categorisation models of metaphor interpretation which have argued that categorisation and comparison statements are fundamentally different (what Chiappe et al. (2003b) call 'the distinct statements view') have also argued that metaphors and similes are understood using the same interpretation mechanisms as their literal counterparts (see, e.g., Glucksberg, 2001, 2008; Chiappe & Kennedy, 2001; Carston, 2002; Wilson & Carston, 2007; Sperber & Wilson, 2008; cf. Grice, 1975; Searle, 1979). We therefore assume that the present research is generally relevant to categorisation models of metaphor interpretation, despite its focus on literal language.

Our study was prompted by the claim, made by proponents of the categorisation view of metaphor interpretation, that comparison statements ('*x* is like a *y*') and categorisation statements ('*x* is a *y*') are incompatible. The purpose of that claim was to argue against the Aristotelian view on metaphor, which holds that metaphors are implicit similes. By contrast, according to categorisation theorists, in 'My lawyer is a shark', for example, the predicate 'shark' expresses a superordinate concept SHARK* which applies to all manner of vicious, predatory creatures, including but not restricted to real sharks. In this view, if metaphors were implicit comparisons, the interpreter of 'My lawyer is a shark' would have to compare my lawyer to SHARK*, which would be as infelicitous as comparing copper to a metal (see Barnden, 2012 for discussion).

Contrary to this claim, we have argued that categorisation statements are compatible with comparison statements. More specifically, '*x* is a *y*' unilaterally entails '*x* is like a *y*'.

Our experimental findings support our claims and show that people are happy to accept that copper is like a metal or that a nuthatch is like a bird (a result that we have replicated in a series of on-line experiments looking at the processing of categorisation and comparison statements; see Xxxxxx, in preparation)[12]. How problematic are these findings for categorisation models of metaphor interpretation? We don't think that our argument undermines the basic insight underlying the categorisation account. In our view, the fundamental tenet of that account is that, in a figurative statement like 'Lawyers are sharks', the predicate 'sharks' does not have its usual lexical meaning. Rather, its ordinary meaning is modulated so as to render it applicable to human beings. Once the contextual meaning of 'sharks' is adapted to the context, 'Lawyers are sharks' is on a par with 'Hammerheads are sharks': both sentences are categorisation statements claiming that the class of entities associated with the subject term fall under the concept associated with the predicate. We consider this to be a valuable insight, and we believe it is true. However, contrary to what categorisation theorists have supposed, we can accept that their central claim is true without having to accept that comparison statements and categorisation statements are fundamentally different.

To prove this point, suppose that categorisation statements were interpreted as follows. (Just suppose: this is *not* a formal proposal. We are fully aware that the following analysis is overly simplistic at best. Its purpose is merely to show that the two claims held by categorisation theorists are in fact independent of one another. It's a proof of concept, not of fact.) Someone utters a sentence of the form 'x is a y' or 'x is like a y'. Let C_x and C_y be the concepts associated with x and y in the context in which the sentence is used. So if the sentence is, 'Hammerheads are sharks', C_x is the standard hammerhead concept and C_y is the standard shark concept. Likewise, if the sentence is, 'Lawyers are

[12] Contrary to the theoretical intuition that 'An apple is like a fruit' is anomalous, the results of various on-line experiments revealed that comparisons to a superordinate are verified equally often and significantly faster than felicitous comparisons such as 'A pear is like an apple'.

like sharks', C_x is the standard lawyer concept and C_y is the standard shark concept. In contrast, if the sentence is, 'Lawyers are sharks', C_x is the standard lawyer concept and C_y is a suitably modulated concept of sharks. (Note that at this point the analysis incorporates what we take to be the key tenet of the categorisation theory.) Suppose, furthermore, that a concept is simply a set of features. Then '*x* is a *y*' is interpreted as meaning that C_y is a subset of C_x (e.g., hammerheads have all properties associated with sharks) and '*x* is like a *y*' is interpreted as meaning that the size of the intersection between C_x and C_y exceeds a context-dependent threshold (e.g., lawyers have some of the properties associated with sharks). Hence, it follows immediately that '*x* is a *y*' and '*x* is like a *y*' are not incompatible and moreover, that the former entails the latter (see also Chiappe & Kennedy, 2001).

As simplistic as it may be, this model proves that the key tenet of the categorisation account of metaphor is consistent with our claim that '*x* is a *y*' and '*x* is like a *y*' are compatible statements. Therefore, contrary to what categorisation theorists have supposed, arguing that metaphors are interpreted as literal categorisation statements does not necessitate the assumption that '*x* is a *y*' and '*x* is like a *y*' are incompatible. It seems to us that, in their zeal to argue against the traditional view that metaphors are implicit similes, categorisation theorists have outreached themselves by defending the much stronger claim that the interpretation of metaphors doesn't involve comparison in any way.

Our argument opens the way for an approach to metaphor which accepts the key tenet of the categorisation view without closing the door on the ancient idea that metaphor involves comparison in some form or other. The simplistic model outlined above suggests one way in which comparison might be involved in the interpretation of metaphors: it could be part and parcel of the meaning of categorisation statements in

general (see Bowdle & Gentner, 2005). There might be other ways, too. For example, comparison might play a role in the construction of figurative meanings, like SHARK* (see Wearing, 2014). In any case, it seems to us that the core intuitions underlying the classical and the categorisation models may well turn out to be compatible.

References

Aristotle (1984). *The complete works of Aristotle: The revised Oxford translation*. J. Barnes (Ed.) Princeton: Princeton University Press.

Barnden, J. A. (2012). Metaphor and simile: Fallacies concerning comparison, ellipsis, and inter-paraphrase. *Metaphor and Symbol, 27*, 265-282.

Bowdle, B. F., & Gentner, D. (2005). The career of metaphor. *Psychological review*, *112*, 193.

Buhrmester, M., Kwang, T., & Gosling, S. D. (2011). Amazon's Mechanical Turk: A new source of inexpensive, yet high-quality, data? *Perspectives on Psychological Science, 6*, 3–5.

Gentner, D., Bowdle, B., Wolff, P., & Boronat, C. (2001). Metaphor is like analogy. In D. Gentner, K. J. Holyoak & B. N. Kokinov (Eds.) (in press). *The analogical mind: Perspectives from cognitive science*. Cambridge, MA: MIT Press.

Carston , R . (2002). *Thoughts and utterances: The pragmatics of explicit communication*. Oxford: Blackwell Publishing.

Carston, R., & Wearing, C. (2011). Metaphor, hyperbole and simile: A pragmatic approach. *Language and Cognition*, *3*, 283–312.

Chiappe, D., & Kennedy, J. (1999). Aptness predicts preference for metaphors or similes, as well as recall bias. *Psychonomic Bulletin and Review*, *6*, 668–676.

Chiappe, D., & Kennedy, J. (2000). Are metaphors elliptical similes? *Journal of Psycholinguistic Research*, *29*, 371–398.

Chiappe, D. L., & Kennedy, J. (2001). Literal bases for metaphor and simile. *Metaphor and Symbol*, *16*, 249–276.

Chiappe, D., Kennedy, J., & Chiappe, P. (2003a). Aptness is more important than comprehensibility in preference for metaphors and similes. *Poetics*, *31*, 51–68.

Chiappe, D., Kennedy, J. M., & Smykowski, T. (2003b). Reversibility, aptness, and the conventionality of metaphors and similes. *Metaphor and Symbol*, *18*, 85–105.

Crump, M. J., McDonnell, J. V., & Gureckis, T. M. (2013). Evaluating Amazon's Mechanical Turk as a tool for experimental behavioral research. *PloS one*, *8*, e57410.

Evans, J. S., Newstead, S.E., & Byrne, R. M. (1993). *Human Reasoning: the Psychology of Deduction*. Hove, East Sussex: Lawrence Erlbaum.
Fogelin, R. J. (1988). *Figuratively speaking*. New Haven, CT: Yale University Press.
Geurts, B. (1998). The mechanisms of denial. *Language, 74,* 274-307.
Geurts, B. (2010). *Quantity Implicatures*. Cambridge University Press.
Geurts, B., & Pouscoulous, N. (2009). No scalar inferences under embedding. In P. Egré and G. Magri (Eds.), *Presuppositions and implicatures*. MIT Working Papers in Linguistics.
Geurts, B. & van Tiel, B. (2013). Embedded scalars. *Semantics and Pragmatics, 6,* 1–37.
Glucksberg, S. (2001). *Understanding figurative language: From metaphors to idioms*. New York, NY: Oxford University Press.
Glucksberg, S. (2003). The psycholinguisitics of metaphor. *Trends in Cognitive Sciences*, *7*, 92–96.
Glucksberg, S. (2008). How metaphors create categories—quickly. In R. W. Gibbs, Jr. (Ed.), *The Cambridge handbook of metaphor and thought* (pp. 67–83). Cambridge, UK: Cambridge University Press.
Glucksberg, S. (2011). Understanding metaphors: The paradox of unlike things compared. In K. Ahmad (Ed.), *Affective computing and sentiment analysis: Emotion, metaphor and terminology* (pp. 1–12). New York, NY: Springer.
Glucksberg, S., & Haught, C. (2006a). On the relation between metaphor and simile: When comparison fails. *Mind and Language*, *21*, 360–378.
Glucksberg, S., & Haught, C. (2006b). Can Florida become like the next Florida? When metaphoric comparisons fail. *Psychological Science*, *17*, 935–938.
Glucksberg, S., & Keysar, B. (1990). Understanding metaphorical comparisons: Beyond similarity. *Psychological Review*, *97*, 3–18.
Glucksberg, S., & Keysar, B. (1993). How metaphors work. In A. Ortony (Ed.), *Metaphor and thought* (2nd edn., pp. 401-424). Cambridge: Cambridge University Press.
Goodman, N.D. & A. Stuhlmüller (2013). Knowledge and implicature: Modeling language understanding as social cognition. *Topics in Cognitive Science* 5: 173–184.
Grice, H. P. (1975). Logic and conversation. In P. Cole & J. Morgan (Eds.), *Syntax and semantics* (Vol. 3, pp. 41–58). New York: Academic Press.
Grice, H. P. (1978). Further notes on logic and conversation. In P. Cole (Ed.), *Syntax and Semantics 9: Pragmatics*, pp. 113–128. New York: Academic Press. Reprinted in and cited from Grice (1989: 41-57).
Grice, H. P. (1989). *Studies in the Way of Words*. Cambridge, Massachusetts: Harvard University Press.
Haught, C. (2013a). A tale of two tropes: How metaphor and simile differ. *Metaphor and Symbol, 28*, 254-274.
Haught, C. (2013b). Spain is not Greece: How metaphors are understood. *Journal of Psycholinguistic Research*. Advance online publication. doi:10.1007/s10936-013-9258-2

Horn, L. R. (1989). *A Natural History of Negation*. Chicago University Press.
Israel, M., Harding, J., & Tobin, V. (2004). On simile. In M. Achard and S. Kemmer (Eds.), *Language, Culture, and Mind* (pp. 123-35). Stanford: CSLI Publications.
Johnson, A. T. (1996). Comprehension of metaphors and similes: A reaction time study. *Metaphor and Symbol*, *11*, 145–159.
Katsos, N. and D. V. Bishop (2011). Pragmatic tolerance: Implications for the acquisition of informativeness and implicature. *Cognition, 120*, 67–81.
Kennedy, J. M., & Chiappe, D. L. (1999). What makes a metaphor stronger than a simile? *Metaphor and Symbol, 14*, 63–69.
Keysar, B. (1989). On the functional equivalence of literal and metaphorical interpretations in discourse. *Journal of Memory and Language, 28*, 375-385.
Kintsch, W. (1974). *The representation of meaning in memory.* New York: Academic Press.
Matsumoto, Y. (1995). The conversational condition on Horn scales. *Linguistics and Philosophy*, *18*, 21–60.
Miller, G. A. (1979). Images and models: Similes and metaphors. In A. Ortony (Ed.), *Metaphor and thought* (pp. 202–250). New York, NY: Cambridge University Press.
O'Donoghue, J. (2009). Is a metaphor (like) a simile? Differences in meaning, effects and processing. *UCL Working Papers in Linguistics*, *21*, 125–149.
Ortony, A. (1979). Beyond literal similarity. *Psychological Review*, *86*, 161–180.
Pierce, R. S. & Chiappe, D. L. (2008). The roles of aptness, conventionality, and working memory in the production of metaphors and similes. *Metaphor and Symbol*, *24*, 1–19.
Rubio-Fernández, P. (2007). Suppression in metaphor interpretation: Differences between meaning selection and meaning construction. *Journal of Semantics, 24*, 345-371.
Schnoebelen, T., & Kuperman, V. (2010). Using Amazon Mechanical Turk for linguistic research. *Psihologija, 43*, 441–464.
Sauerland, U. (2004). Scalar implicatures in complex sentences. *Linguistics and Philosophy*, *27*, 367-391.
Searle, J. (1979). Metaphor. In A. Ortony (Ed.), *Metaphor and thought* (pp. 92–123). New York, NY: Cambridge University Press.
Soames, S. (1982). How presuppositions are inherited: a solution to the projection problem. *Linguistic Inquiry, 13*, 483–545.
Sperber, D., & Wilson, D. (2008). A deflationary account of metaphor. In R.W. Gibbs, Jr. (Ed.), *The Cambridge handbook of metaphor and thought* (pp. 84–105). Cambridge, UK: Cambridge University Press.
Sprouse, J. (2011). A validation of Amazon Mechanical Turk for the collection of acceptability judgments in linguistic theory. *Behavior Research Methods*, *43*, 155–167.
Stern, J. (2000). *Metaphor in context*. Cambridge, MA and London, UK: Bradford Books, MIT Press.

Utsumi, A. (2007). Interpretive diversity explains metaphor-simile distinction. *Metaphor and Symbol*, *22*, 291–312.
van Rooij, R. and Schulz, K. (2004). Exhaustive interpretation of complex sentences. *Journal of Logic, Language and Information, 13,* 491–519.
van Tiel, B. (2014a). Embedded scalars and typicality. *Journal of Semantics, 31*, 147–177.
van Tiel, B. (2014b). *Quantity matters: Implicatures, typicality and truth*. Doctoral dissertation, University of Nijmegen.
Wearing, C. (2014). Interpreting novel metaphors. *International Review of Pragmatics. 6,* 78-102.
Wilson, D. & Carston, R. (2007). A unitary approach to lexical pragmatics: Relevance, inference and ad hoc concepts. In N. Burton-Roberts (Ed.), *Advances in Pragmatics* (pp. 230-260). Basingstoke: Palgrave Macmillan.
Zharikov, S. & Gentner, D. (2002). *Why do metaphors seem deeper than similes?* In Proceedings of the 24th Annual Meeting of the Cognitive Science Society (pp. 976–981). Fairfax, VA: Cognitive Science Society.
Xxxxx (in preparation). The relatedness effect in sentence verification: Are categorisaton and comparison statements related?

Appendix: Experimental Report

Experiment 1a

Method

Participants

50 participants were recruited through Mechanical Turk.

Materials and procedure

We constructed a total of 108 slides including either a comparison or a categorisation statement (e.g., 'This one is like a moirk'). The statement referred to a figure that was presented on the same slide (e.g., a triangle) and included a made-up category that was also defined on the same slide. The definitions consisted of three properties of each figure: shape, colour and border type (e.g., 'Moirk: circle, blue, thick border'). The figure

on the slide could either share no property with the defined category (Similarity Level 0), one property (Similarity Level 1), two properties (Similarity Level 2) or all three properties (Similarity Level 3).

The critical items consisted of 40 pairs of slides, each pair including a comparison statement and a categorisation statement and 10 pairs corresponding with each of the 4 possible degrees of similarity (i.e. Level 0-3). An extra 20 filler items included a true categorisation statement (Level 3) and another 8 items, one of each type, were used as warm-up trials. The warm-up trails were presented in the same random order to all participants, while the critical and filler items were randomized individually.

Participants were asked to verify the statement that appeared at the top of each slide in relation to the figure and the definition that were presented underneath the statement. Participants were given a TRUE/ FALSE choice to respond.

Results

The mean proportions of TRUE responses in each condition are plotted in Figure 2. The overall pattern of results suggests that participants didn't derive a non-*y* inference in interpreting the comparison statements.

We fitted a logistic mixed-effects model, positing fixed effects of Statement Type and Similarity Level, and random effects of Participant and Item, as well as a random slope of Statement Type by Participant. (Models with additional random slopes did not converge.) This disclosed significant main effects of Statement Type and Similarity Level, and a significant interaction ($p < 0.001$, model comparison).

Follow-up pairwise comparisons at different similarity levels were implemented using logistic mixed-effects models, positing a fixed effect of Statement Type and random effects of Participant and Item, as well as a random slope of Statement Type by

Participant. There was a significant main effect of Statement Type in the Level 2 condition (β = 3.22, SE = 0.67, Z = 4.80, $p < 0.001$), but no significant main effect in the Level 1 condition (β = 0.70, SE = 3.34, Z = 0.21, p = 0.833) or the Level 3 condition (β = 13.3, SE = 14.3, Z = 0.93, p = 0.352). The significant effect remains significant ($p < 0.001$) when corrected for multiple comparisons.

Looking at individual performances, 25 of the 50 people who took part in the task adopted the categorisation strategy by default. That is, they responded TRUE in all cases of maximal similarity, responding FALSE in all Similarity Level 0-2 trials regardless of the type of statement. Removing those 25 participants from the analyses did not change the overall pattern of results greatly (see Figure 9 below). There was a 9% increase in the proportion of TRUE responses in the Level 1 condition and a 37% increase in the Level 2 condition, but the mean proportion of TRUE responses in the Level 3 condition was still .90 (decreasing only 5% from the overall analyses). More specifically, only 2 participants systematically responded FALSE in the Comparison/ Level 3 condition.

Experiment 1b

Given that half of the participants in Experiment 1a applied the same strategy in the Categorisation and Comparison conditions (i.e., responded TRUE only in cases of maximal similarity), Experiment 1b tried to determine whether the results of the Comparison/ Level 3 condition may have been skewed. More specifically, we wanted to determine whether participants may derive a non-*y* inference in interpreting comparison statements if they are not presented with categorisation statements in the same task.

Method

Participants

20 participants were recruited through Mechanical Turk.

Materials and procedure

The materials were those used in Experiment 1a for the Comparison condition. The procedure was the same as in the first experiment.

Results

One participant was eliminated because he had responded randomly. The mean proportions of TRUE responses in each condition are plotted in Figure 9. The overall pattern of results suggests that participants didn't derive a non-*y* inference in interpreting the comparison statements in Experiment 1b, in line with what was observed in Experiment 1a.

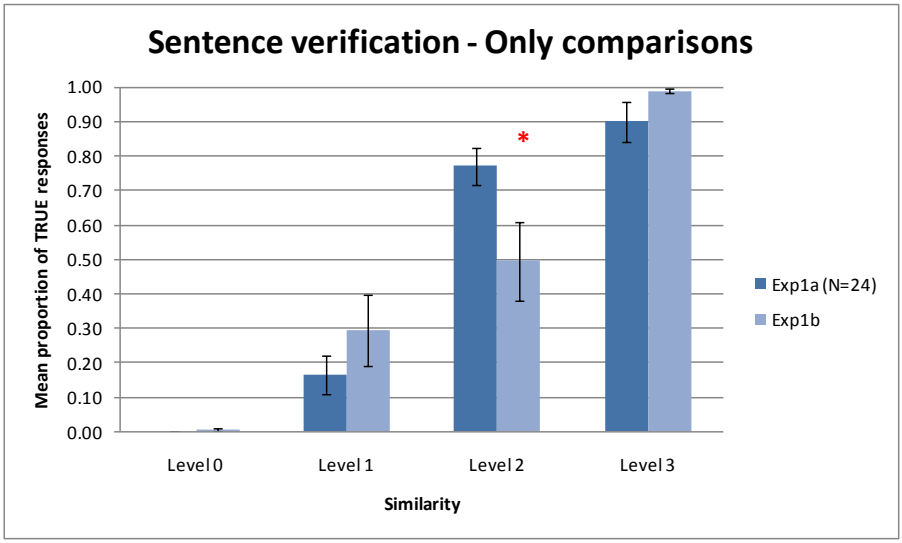


Figure 9: Mean proportions of TRUE responses to comparison statements from those participants in Experiment 1a who didn't adopt the categorisation strategy by default and from Experiment 1b, which included only comparison statements (SE bars; asterisk $p < .03$).

Rather than increasing the proportion of TRUE responses in the Similarity Level 1 and Level 2 conditions, presenting participants only with comparison statements resulted in comparable agreement rates in the Level 1 condition and significantly lower agreement rates in the Level 2 condition. More importantly, in the Similarity Level 3 condition, the results were comparable between the two experiments. In fact, participants were more prone to agreeing with the comparison statements when they were not presented with categorisation statements in the same task and none of the participants in Experiment 1b systematically rejected the comparison statements in the maximal-similarity condition (while 2 participants had done so in Experiment 1a). These results therefore suggest that the results of Experiment 1a were reliable and not an artefact of presenting participants with comparison and categorisation statements in the same task.

Experiment 2

Method

Participants

20 participants were recruited through Mechanical Turk.

Materials and procedure

We constructed 80 slides including a picture of an animal and a description of the animal in a bubble (e.g., 'This one is a wild animal' referring to a tiger). The descriptions made up 40 pairs of categorisation and comparison statements about 10 well-known animals. We used 4 different types of categories: (1) superordinates (e.g., 'wild animal' for a tiger), (2) same category (e.g., 'tiger' for a tiger), (3) similar category (e.g., 'lion' for a tiger) and (4) dissimilar category (e.g., 'bear' for a tiger).

The materials were presented in the same quasi-random order to all participants. In order to avoid possible issues with the identification of the animals in the pictures, the categorisation version of each Same-Category item was presented before the comparison version (i.e., participants had to agree to 'This one is a tiger' before they had to decide on 'This one is like a tiger').

Participants were asked to verify a series of facts about 10 well-known animals that were presented in pictures. Participants were given a TRUE/ FALSE choice to respond.

Results

The mean proportions of TRUE responses for each condition are plotted in Figure 4. The overall pattern of results suggests that participants didn't derive a non-*y* inference in interpreting the comparison statements relative to what is observed for the categorisation statements.

We fitted a logistic mixed-effects model, positing fixed effects of Statement Type and Category Type, and random effects of Participant and Item, as well as a random slope of Statement Type by Participant. (Models with additional random slopes did not converge.) This disclosed significant main effects of Statement Type and Category Type, and a significant interaction ($p < 0.001$, model comparison).

Follow-up pairwise comparisons for different category types were implemented using logistic mixed-effects models, positing a fixed effect of Statement Type and random effects of Participant and Item, as well as a random slope of Statement Type by Participant. There was a significant main effect of Statement Type in the Similar condition ($\beta = 3.73$, SE = 1.04, $Z = 3.59$, $p < 0.001$), which remains significant when corrected for multiple comparisons ($p < 0.001$), but no significant main effect in the

Dissimilar condition (β = 1.68, SE = 1.73, Z = 0.97, p = 0.33), Same condition (β = 4.67, SE = 4.90, Z = 0.95, p = 0.34) or Superordinate condition (β = 3.49, SE = 2.54, Z = 1.37, p = 0.171).

Experiment 3

Method

Participants

25 participants were recruited through Mechanical Turk.

Materials and procedure

We constructed 42 different items of the form

"John says: My mother is like a nurse.

Would you conclude from this that, according to John, his mother is not a nurse?"

The critical items consisted of 24 items evenly distributed in 4 conditions: true conclusion according to common knowledge (T/C); false conclusion according to common knowledge (F/C); true conclusion according to common knowledge if the conclusion is interpreted literally (T(lit.)/C), and possible conclusion according to the speaker's private knowledge (P/SP). See Table 1 for an example of each condition.

The statements in the T(lit.)/C condition were similes and the conclusions negated metaphors. Taken literally, the conclusions were obviously true since the metaphors included a category violation (e.g., 'His ideas are like diamonds' –> 'His ideas are not diamonds'). The similes/ metaphors were relatively conventional.

We used 12 filler items in the standard form, and 6 control items in which the statement was a relatively conventional metaphor and participants had to decide whether the

corresponding simile followed from the statement (e.g., 'My lawyer is a shark' –> 'My lawyer is like a shark').

The materials were presented in the same random order to all participants. Participants were asked to decide whether, according to the person making the statement, a certain conclusion followed from what they had said. Participants were given a YES/ NO choice to respond.

Results

The mean proportions of TRUE responses in each condition are plotted in Figure 7. The results show different agreement rates in the various conditions depending on the hearer's knowledge.

We fitted a logistic mixed-effects model, positing fixed effects of Condition and random effects of Participant and Item, as well as a random slope of Condition by Participant. This disclosed a significant main effect of Condition ($p < 0.01$, model comparison).

Follow-up pairwise comparisons were implemented by using the same model over subsets of the data. These models disclosed significant differences between the T/C condition and each of the other three conditions (F/C: $\beta = 5.17$, SE = 1.33, $Z = 3.88$, $p < 0.001$; T(lit.)/C: $\beta = 6.73$, SE = 3.05, $Z = 2.21$, $p < 0.05$; P/SP: $\beta = 5.28$, SE = 1.69, $Z = 3.13$, $p < 0.001$). Corrected for multiple comparisons, the T/C to F/C comparison is significant with $p < 0.001$, the T/C to T(lit.)/C comparison is marginally significant ($p = 0.082$) and the T/C to P/SP comparison is significant with $p < 0.01$. None of the other pairwise comparisons showed significant differences (F/C vs. T(lit.)/C: $\beta = 0.67$, SE = 1.08, $Z = 0.62$, $p = 0.535$; F/C vs. P/SP: $\beta = 0.356$, SE = 0.907, $Z = 0.39$, $p = 0.695$; T(lit.)/C vs. P/SP: $\beta = 0.072$, SE = 0.616, $Z = 0.12$, $p = 0.907$).

Experiment 4

Method

Participants

25 participants were recruited through Mechanical Turk.

Materials and procedure

We constructed 54 slides each including a picture of an animal, fruit, or vegetable in the center. Underneath the picture was a short description, which could be in categorisation form (e.g., 'Robins are birds') or in comparison form (e.g., 'Chickens are like a farm animals').

The critical items were 18 pairs of categorisation and comparison statements about 18 well-known animals. 3 types of categories were used in the statements: (1) basic level (e.g., 'Labradors are dogs'), (2) superordinate (e.g., 'Sharks are predators') and (3) name repeated (i.e., the category was already mentioned in the name of the animal; e.g., 'Grizzly bears are bears').

18 comparison statements were used as fillers, 9 depicting animals and 9 depicting fruits or vegetables. The filler items included two similar animals, fruits or vegetables (e.g., 'Wild boars are like pigs'/ 'Shallots are like onions'). The two-animal fillers were used as a baseline for the critical comparison statements. The materials were presented in the same random order to all participants.

Participants were asked rate a series of descriptions of animals, vegetables and fruits on a 1-7 scale ranging from 1 = Completely unacceptable to 7 = Perfectly acceptable.

Results

The mean appropriatness ratings for each condition are plotted in Figure 6. The overall pattern of results suggests that categorisation statements were preferred over comparison statements in all conditions.

As participants were generally consistent across items within each condition, we consider their means for each condition (i.e. we consider each participant to give rise to one data point per condition). Paired t-tests reveal highly significant differences between categorisation and comparison in each condition (all $p < 0.001$) and between each comparison condition and the control (all $p < .002$). However, as the data are not normally distributed, we also report the results of a non-parametric statistical test, namely the sign test. There is a highly significant preference for categorisation in all three conditions (21/25 participants in the Basic-Level condition, 21/25 participants in the Repetition condition, and 22/25 participants in the Superordinate condition: all $p < 0.001$). There is also a highly significant preference for the comparisons in the control condition than in each of the other comparison conditions (19/25 participants in the Basic-Level condition, 19/25 participants in the Repetition condition, and 20/25 participants in the Superordinate condition: all $p < 0.008$).

Experiment 5

Method

Participants

200 participants were recruited through Mechanical Turk.

Materials and procedure

All participants were randomly presented with one of 4 types of narrative, 2 control and 2 critical conditions. For the actual narrative, see the main text. At the end of the

narrative, participants were given a 1-7 Likert scale to indicate their interpretation of the final statement, with 1 meaning 'Definitely disagree' and 7 meaning 'Definitely agree'.

Results

The mean ratings for each condition are plotted in Figure 8. In the Critical/ Partial knowledge condition, the mean rating was 4.94 (SD 1.64), with 10 participants giving the highest possible rating of 7. In the corresponding Full knowledge condition, the mean rating was 6.08 (SD 1.57), with 29 participants giving a maximum rating of 7. An unpaired t-test shows the difference in rating to be highly significant ($t = 3.55$, df = 98, $p < 0.001$). As the ratings are not normally distributed, we also consider the proportion of maximum ratings given in each condition: this exhibits a highly significant difference (Fisher's exact test, $p < 0.001$).

By comparison, in the Control/ Partial knowledge condition, the mean rating was 5.50 (SD 1.91), with 20 participants giving the rating of 7. In the corresponding Full knowledge condition, the mean rating was 6.32 (SD 1.32), with 31 participants giving the rating of 7. An unpaired t-test shows the difference in rating to be significant ($t = 2.50$, df = 98, $p < 0.016$), and the proportion of maximal ratings also differs significantly (Fisher's exact test, $p < 0.05$).